UMUTeam at SemEval-2023 Task 3: Multilingual transformer-based model for detecting the Genre, the Framing, and the Persuasion Techniques in Online News

Ronghao Pan[1], **José Antonio García-Díaz**[1],
Miguel Ángel Rodríguez-García[2], **Rafael Valencia-García**[1]
[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan@um.es, joseantonio.garcia8, valencia}@um.es
[2]Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos,
28933 Madrid, Spain miguel.rodriguez@urjc.es

Abstract

In this manuscript, we describe the participation of the UMUTeam in SemEval-2023 Task 3, a shared task on detecting different aspects of news articles and other web documents, such as document category, framing dimensions, and persuasion technique in a multilingual setup. The task has been organized into three related subtasks, and we have been involved in the first two. Our approach is based on a fine-tuned multilingual transformer-based model that uses the dataset of all languages at once and a sentence transformer model to extract the most relevant chunk of a text for subtasks 1 and 2. The input data was truncated to 200 tokens with 50 overlaps using the sentence-transformer model to obtain the subset of text most related to the articles' titles. Our system has performed good results in subtask 1 in most languages, and in some cases, such as French and German, we have archived first place in the official leader board. As for task 2, our system has also performed very well in all languages, ranking in all the top 10.

1 Introduction

This working note describes the participation of the UMUTeam in the SemEval 2023 shared task 3 (Piskorski et al., 2023), focused on identifying different aspects of news articles, including document category, framing dimensions, and persuasion technique in a multilingual setup.

Online media monitoring in multiple languages is an essential capability for organizations and companies that analyze and report on specific topical issues in different countries. However, due to the large amount of data to be processed and the interest in comparing in depth how topics of interest are treated and presented in different countries and media, the automation of media analysis processes is necessary. Automating media analysis processes can help reduce time and effort while providing a more complete and detailed overview of media coverage in different languages and geographies. It can also help identify trends, patterns, and deviations that would otherwise be difficult to detect. Therefore, this task focuses on recognizing different complementary aspects of a persuasive text, such as genre and persuasive techniques used to influence the reader.

Advances in deep learning have enabled the development of very sophisticated and deep machine learning models. These include transformer-based models, a neural network architecture that has a weighty impact on NLP, making them suitable for tasks such as text classification, automatic translation, and sentiment analysis (Bozinovski, 2020). In addition, they are trained on large datasets and are able to learn complex features in different languages. Multilingual transformer-based models are a type of model that has been trained in several languages and therefore have the ability to deal with the linguistic task involving more than one language. Thus, our approach for subtasks 1 and 2 is to fine-tune a multilingual model called XLM-RoBERTa using a dataset of all languages at once and with particular pre-processing techniques that will be explained in more detail in Section 3. Our system has achieved good results in subtask 1 in most languages, such as French and German, in which we have achieved first place in the official ranking. As for task 2, our system has also performed very well in all languages, ranking in all the top 10.

The remainder of this paper is organized as follows. Section 2 provides a summary of important details about the task setup. Section 3 offers an overview of our system for two subtasks. Section 4 presents the specific details of our systems. Section 5 discusses the results of the experiments, and finally, the conclusions are shown in Section 6

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 609–615

2 Background

The SemEval-2023 task 3 challenge consists mainly of the following sub-task:

1. **News genre categorization**: Identify whether a news article is an opinion piece, aims at objective news reporting, or is a satire piece. It is a single-label task at the article level.

2. **Framing detection**: Identify the frames uses in the article. It is a multi-label task at the article level.

3. **Persuasion techniques detection**: determine the persuasion techniques in each paragraph. It is a multi-label task at the paragraph level.

We have only used the dataset given by the organizers, which consists of a set of news articles in 6 languages (English, French, German, Italian, Polish, and Russian). The dataset has been compiled from 2020 to mid-2022, revolving around a fixed set of widely discussed topics, such as COVID-19, global climate change, abortion, migration, preparations leading up to the Russian-Ukrainian war and events related to and triggered by it, as well some country-specific local events, such as elections, etc. The media selection has covered both mainstream media in each country, such as alternative news and web portals. Some news aggregation engines were used, such as Google News or Europe Media Monitor (EMM), a large-scale, multilingual, near-real-time news aggregation and analysis engine. Moreover, online services were used, such as NewsGuard and MediaBiasFactCheck, which rank sources according to their likelihood of spreading misinformation or disinformation.

It is worth mentioning that task's organizers divided the corpus of all languages into three parts: training, development, and testing. However, the size of the training and development datasets is relatively small, as shown in Table 1, so we have decided to combine the training and development data from all languages to fine-tune and optimize the hyperparameters of the XLM-RoBERTa model.

3 System Overview

For solving the subtask 1 and 2 proposed, we build a system, which architecture is depicted in Figure 1. In a nutshell, our system works as follows. First, the training and evaluation set of all languages was merged, and a preprocessing stage was performed to clean the dataset. Second, a Sentence-Transformers model was used to extract the part of the news item most related to the title. Third, we fine-tuned the XLM-RoBERTa model for different classification tasks. Finally, we evaluated the model with test datasets for each language.

3.1 Preprocessing Stage

Our preprocessing stage included the following steps: (1) removing social media languages, such as hyperlinks, hashtags, or mentions, (2) removing the character "*", and (3) removing all email links.

3.2 Sentence Transformer

When analyzing the texts of news articles it has been detected that most of them have very long texts, reaching in some cases up to 7655 tokens at most, which makes it difficult to learn the models through the document features. To reduce the text size and extract only the most important part of the news, a model called Sentence-Transformers (Reimers and Gurevych, 2020) has been used for paraphrase mining. Paraphrase mining consists of identifying the sentences or pairs of sentences with high similarity. Thus, for subtasks 1 and 2, we divided the text into 200 words fragments with 50 overlapping tokens and used Sentence-Transformers to identify the fragment with the highest relationship to the news title. For our approach, we have used a multilingual sentence transformer model called *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2*[1].

3.3 Classification Model

Our approach for subtasks 1 and 2 is based on fine-tuning XLM-RoBERTa, and the processes to be performed are as follows.

3.3.1 Tokenization

The main feature of the transformer networks is their self-attention mechanism, whereby each word in the input is able to learn what relationship it has with others (Yi and Tao, 2019). All pre-trained models require the input data to be pre-processed through the process of tokenization, which consists of decomposing a large entity into smaller components called *tokens*. In this case, XLM-RoBERTa uses subword tokenization with the Byte-Pair Encoding (BPE) algorithm.

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Table 1: Corpus statistics for subtask 1

Split	English	French	German	Italian	Polish	Russian	Total
Train	433	157	132	226	144	142	1234
Dev	83	54	45	77	50	49	358
Test	54	50	50	61	47	72	334

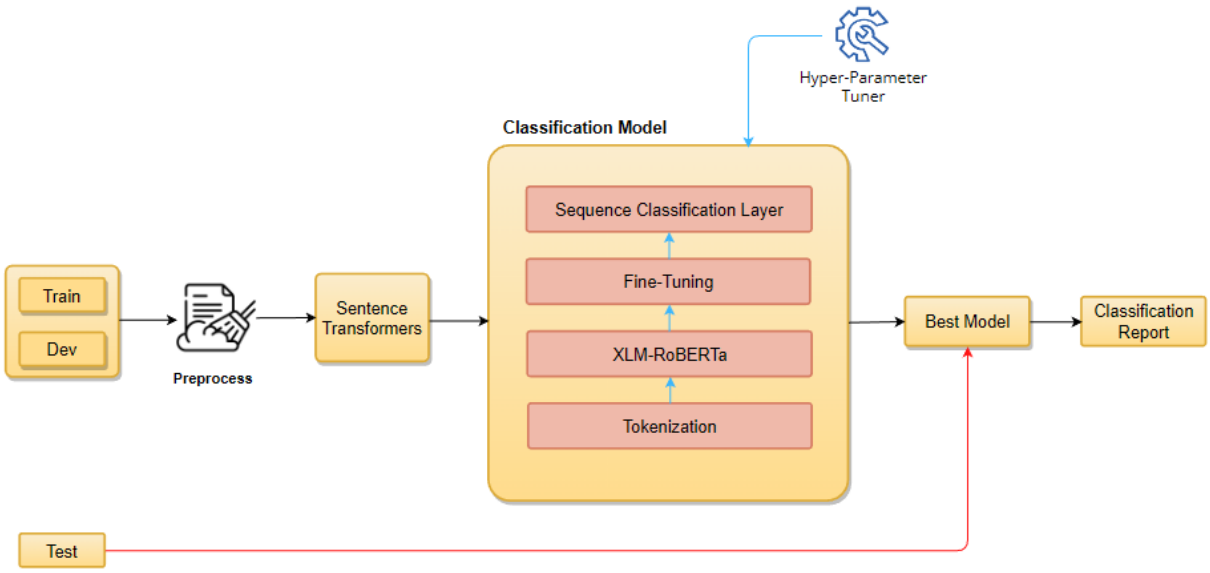


Figure 1: Overall system architecture.

3.3.2 XLM-RoBERTa

XLM-RoBERTa[2] is a multilingual version of RoBERTa pre-trained with 2.5 TB of filtered CommonCrawl data containing 100 languages. The model is able to use an internal representation of 100 languages to extract features useful for tasks such as classification, sentiment analysis and others (Conneau et al., 2020).

3.3.3 Fine-tuning

Fine-tuning is a deep learning model training process in which a pre-trained model is taken as a basis and additionally trained on a specific task with a smaller dataset. In this way, the prior knowledge of the pre-trained model can be exploited and its behavior adjusted. In this case, in the absence of sufficient training data, as shown in Table 1, this process improves the overall performance of the model compared to training from scratch with the corpus provided by the organizers.

For subtasks 1 and 2, we have used the structure (see Figure 1), since for both subtasks we add a sequence classification layer on top of the pre-trained model and train this layer for the detection of the genre of the news item and the frames used. The main difference is that subtask 2 is a multi-label classification, so a piece of news can use more than one frame, which implies that the last sequence classification layer is of a multi-label type.

4 Experimental Setup

Finally, the dataset for training has a total of 1234 news items and 258 for development for both subtask 1 and subtask 2. To obtain the classification model for each task, we conducted a hyperparameter optimization stage using RayTune ((Bergstra et al., 2013)) with a Tree of Parzen Estimators (TPE) to select the best combination of the hyperparameters over 10 trials. The hyperparameters evaluated, and their interval range are: (1) weight decay (between 0 and 0.3), (2) training batch size ([8,16]), (3) number of training epochs ([1-20]), and (4) learning rate (between 1e-5 and 5e-5). For subtasks 1 and 2, the best subset of hyperparameters for XLM-RoBERTa is 15 epochs, 0.01 in weight decay, a batch training of 8, and a learning rate of 1e-5. The input size of the text is also an

[2] https://huggingface.co/xlm-roberta-base

important factor that significantly affects the performance of our system, so we tested taking fragments of different sizes with the phrase transformer model, such as 200, 300, 400, and 500 tokens and found that the best performance was obtained with fragments of 200 tokens with 50 overlap as input.

The primary evaluation metric used in the task is the F1 scores. For sub-task 1, the Macro-F1 is the official evaluation measure. For sub-task 2, the evaluation metric is Micro-F1, in this case, being a multi-label classification, the Seqeval[3] framework has been used as the evaluation framework when performing hyperparameter optimization and Micro-F1 comparisons.

5 Results

Our system was evaluated using the validation split and trained with the set of all languages at once.

5.1 Subtask 1

The results for subtask 1 on the validation dataset set are depicted in Table 2. As can be seen in Table 2, the system is very limited for the classification English news category, with a Macro F1-score (M-F1) of 0.29408 in the validation set, and that is because the dataset is unbalanced and there are very few examples of satire and reporting type news in the training set, as shown in Table 3. However, in other languages, having a less unbalanced dataset in the case of French, German, and Polish, better results have been obtained. A normalized confusion matrix with truth labels of the best language models on the validation dataset has been utilized to identify scenarios where models generate inaccurate predictions (see Figure 2). This matrix is presented in the form of a chart showing the distribution of a model's predictions compared to the actual truth labels of the data. As shown in Figure 2, our system often confuses the reporting and opinion type because it does not have enough cases of reporting in the training set. Concerning satire, there is often confusion about opinion-type news.

Table 4 depicts the official results for subtask 1. As can be observed, our submission outperforms all the proposed baselines, achieving in all the top 10, and in French and German, we have reached the top 1. Another advantage of our approach is that being a multilingual model and not having training data for Spanish, Greek, and Georgian, the model performs quite well in a zero-shot learning scenario, as in Table 4. It is possible, as we have taken advantage of prior knowledge of the XLM-RoBERTa model and fine-tuned it for this classification task.

[3] https://github.com/chakki-works/seqeval

Table 2: Result for the subtask 1, reporting the F1-score of the opinion, reporting and satire, and the macro F1-score (M-F1).

Language	Opinion	Reporting	Satire	M-F1
En	0.3855	0.3429	0.1539	0.2941
Fr	0.8831	0.7200	0.6667	0.7566
Ge	0.8214	0.5714	0.7692	0.7207
It	0.8976	0.4348	0.5000	0.6108
Po	0.9041	0.7059	0.8000	0.8033
Ru	0.8437	0.6897	0.4000	0.6445

Table 3: Distribution of the training dataset by news categories and languages.

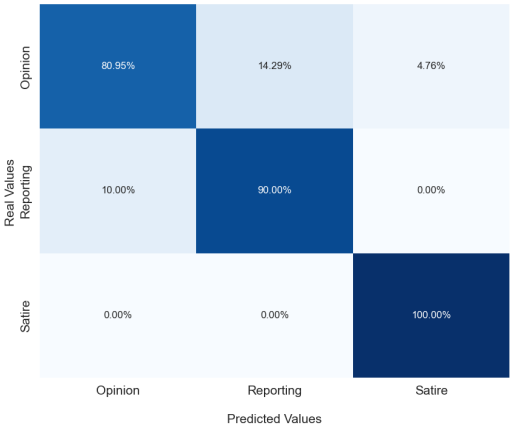
Language	Opinion	Reporting	Satire
En	382	41	10
Fr	103	43	11
Ge	86	27	19
It	174	44	8
Po	104	25	15
Ru	93	41	8

Table 4: Official results for the subtask 1, reporting the rank, the macro F1-score the baseline, the macro F1-score (M-F1), and the micro F1-score (m-F1).

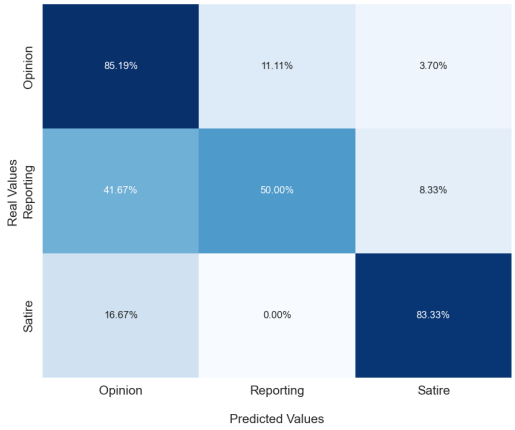
Language	Rank	Baseline	M-F1	m-F1
En	11	0.2880	0.4134	0.5926
Fr	1	0.5681	0.8355	0.8800
Ge	1	0.6296	0.8195	0.8200
It	6	0.3894	0.5534	0.7541
Po	5	0.4896	0.6643	0.8085
Ru	4	0.3983	0.6454	0.6806
Es	6	0.1538	0.4375	0.5000
Gr	2	0.1705	0.7670	0.7969
Ka	6	0.2564	0.5820	0.8620

5.2 Subtask 2

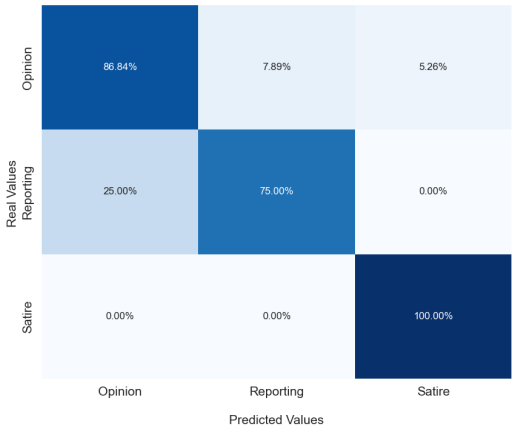
Subtask 2 is a multi-label classification problem in which, given a news article, the model must identify the frames used in the article. A frame is a perspective under which an issue or a piece of news is presented. In this case, the organizers have followed the frames defined by (Card et al., 2015), and there are a total of 14 frames: Economic, Capacity and resources, Morality, Fairness and equality, Legality, constitutionality and jurisprudence, Policy



(a) French



(b) German



(c) Polish

Figure 2: Confusion matrix of the best models according to language.

prescription and evaluation, Crime and punishment, Security and defense, Health and safety, Quality of life, Cultural identity, Public opinion, Political, External regulation, and reputation.

Table 5 shows the results of the evaluation with the validation set, and it can be seen there is quite a large difference between the micro F1-score (m-F1) and the macro F1-score (M-F1) and that is because the dataset is unbalanced. The micro F1-score counts the total number of correct and incorrect predictions, while the macro F1 calculates the precision and the recall of each class and then averages them, so it takes into account the number of examples in each class. For this task, the evaluation metric is micro F1, and our approach has performed better on the validation dataset than the baseline in all languages. In this case, the models work really well for English and Polish, with a micro F1 of 0.73514 and 0.65342.

Finally, the official results for the subtask 2 are depicted in Table 6. It can be seen that our system has achieved a top 10 ranking in all languages and outperforms all the proposed baselines. Another advantage of our approach is that, being a multilingual model, it obtained good results in languages not included into the training set, such as Spanish, Greek, and Georgian (zero-shot learning scenario), due to the transfer learning approach.

6 Conclusion

In this working note, we have described the participation of the UMUTeam in the shared task 3 of SemEval 2023. In this shared task, the participants were required to detect different aspects of news articles and other web documents, such as document categories, framing dimensions, and per-

Table 5: Result for the subtask 2, reporting the micro F1-score (m-F1), and the macro F1-score (M-F1).

Language	m-F1	M-F1
En	0.7351	0.4159
Fr	0.5524	0.4687
Ge	0.5990	0.4935
It	0.5880	0.4989
Po	0.6534	0.5496
Ru	0.4718	0.3495

Table 6: Official results for the subtask 2, reporting the rank, the micro F1-score of the baseline, the macro F1-score (M-F1), and the micro F1-score (m-F1)

Language	Rank	Baseline	M-F1	m-F1
En	9	0.3499	0.4152	0.5083
Fr	8	0.3285	0.4773	0.4382
Ge	8	0.4871	0.5654	0.6138
It	6	0.4856	0.4467	0.5763
Po	4	0.5938	0.5931	0.6418
Ru	8	0.2295	0.2884	0.3849
Es	2	0.1200	0.4654	0.5584
Gr	3	0.3454	0.4036	0.5341
Ka	5	0.2597	0.4106	0.5294

suasion techniques in a multilingual setup. Our system has achieved good results in subtask 1 in most languages, such as French and German, and we have archived first place in the official ranking. As for task 2, our system has also performed well in all languages, ranking in all the top 10.

In the future, we hope to explore other multilingual models and fine-tuning a Sentence Transformer model using the news dataset to improve the model performance and perform an ablation study on fragment size to analyse deeper the impact of this parameter. Besides, we want to incorporate and to extend the Spanish split of the evaluation dataset to the news related to the financial domain (García-Díaz et al., 2023), hate-speech detection (García-Díaz et al., 2022a) and political ideology (García-Díaz et al., 2022b).

7 Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

References

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA. PMLR.

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)*, 44(3).

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023. Smart analysis of economics sentiment in spanish based on linguistic features and transformers. *IEEE Access*.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022a. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

José Antonio García-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín Valdivia, Francisco García-Sánchez, L Alfonso Ureña-López, and Rafael Valencia-García. 2022b. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguaje Natural*, 69:265–272.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jiangyan Yi and Jianhua Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274.